

Minería de datos y Patrones

Version 2025-I

Fundamentos Matemáticos para Minería de Datos y Machine Learning

Dr. José Ramón Iglesias

DSP-ASIC BUILDER GROUP

Director Semillero TRIAC

Ingeniería Electronica

Universidad Popular del Cesar

Introducción

En minería de datos y machine learning necesitamos dominar conceptos y herramientas provenientes de las siguientes áreas de las matemáticas.

- **Álgebra Lineal**
- **Optimización**
- **Probabilidades**
- **Inferencia Estadística**

A continuación se muestra un repaso de los conceptos más importantes de estas disciplinas.

Álgebra Lineal

Álgebra Lineal

En ML usamos las siguientes representaciones algebraicas para nuestros objetos y sus atributos: **escalares**, **vectores** y **matrices**.

- **Escalares**: un escalar es simplemente un número como 7.56. El valor de un atributo numérico para un objeto se representa por un escalar.
- **Vectores**: un vector es un arreglo ordenado de escalares $\mathbf{x} = [x_1, x_2, \dots, x_n]$ donde x_i es el i -ésimo elemento de \mathbf{x} .
 - a. En ML un objeto de n **atributos** numéricos puede ser representado por un vector de n dimensiones.
 - b. El **producto punto** entre dos vectores $\mathbf{a} \cdot \mathbf{b}$ es la suma de la multiplicación de todos sus elementos:

$$(a_1, a_2, a_3, \dots, a_n) \cdot (b_1, b_2, b_3, \dots, b_n) = a_1 b_1 + a_2 b_2 + \dots + a_n b_n = \sum a_i \cdot b_i$$

Norma y Distancia Euclidiana

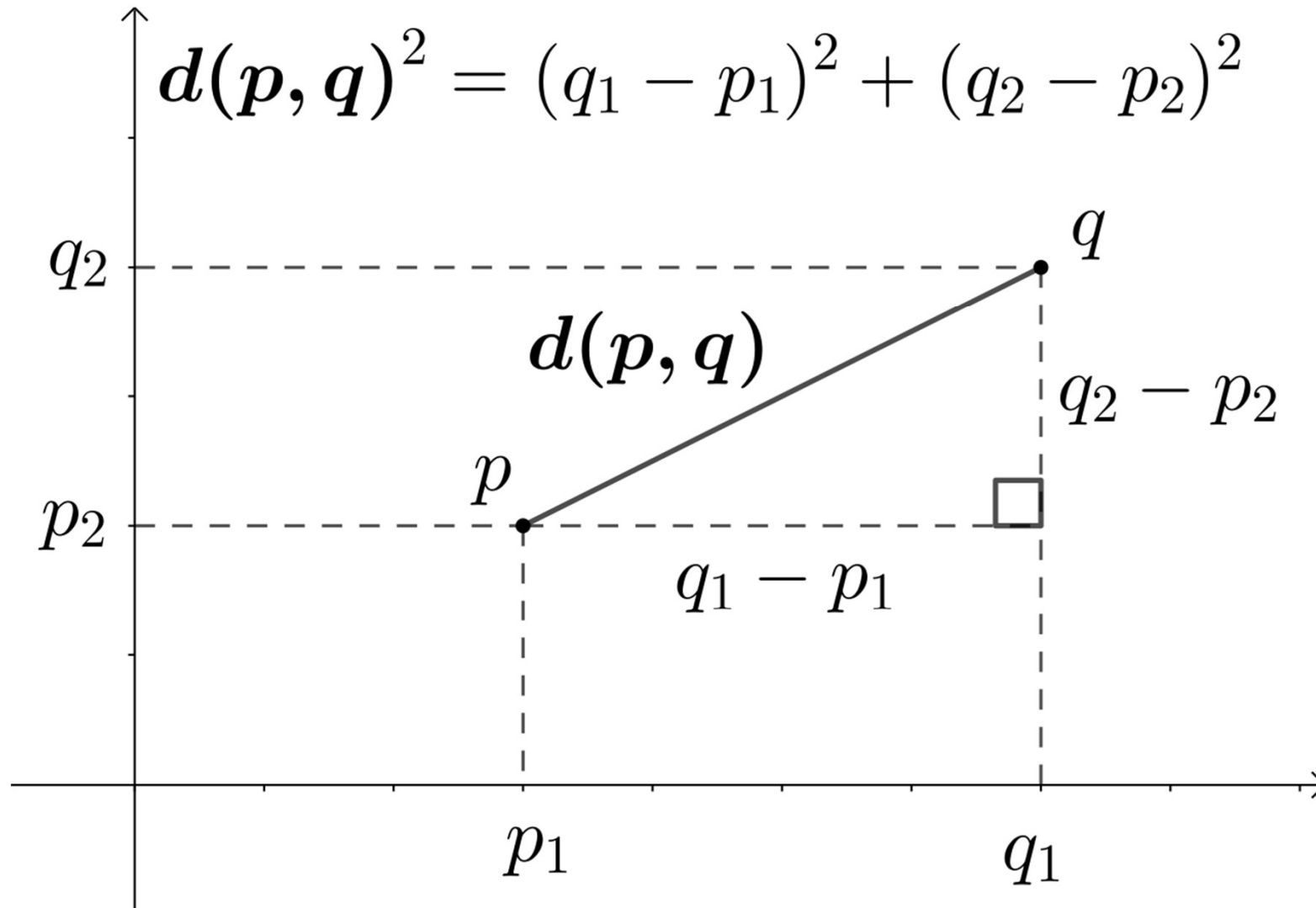
- La norma euclidiana de un vector $\|\mathbf{v}\|_2$ es el largo del vector en un espacio euclidiano (piensen en pitágoras).

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2} = \sqrt{\sum_{i=1}^n v_i^2}$$

- Luego, la distancia euclidiana nos permite calcular qué tan lejos están dos vectores \mathbf{x} e \mathbf{y} .

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2},$$

Distancia Euclidiana



Distancia Euclidiana en \mathbb{R}^2 . Fuente: Wikipedia

Matrices

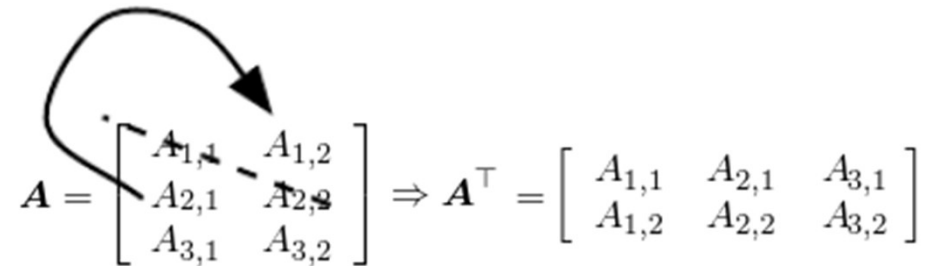
- Una matriz es un arreglo de dos dimensiones, entonces cada elemento se identifica por dos índices en vez de uno.
- Ejemplo: sea A una matriz de dos filas y dos columnas ($A \in \mathbb{R}^{2 \times 2}$)

$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- Un dataset de n objetos y m atributos numéricos se puede representar por una matriz de $n \times m$.
- El i -ésimo ejemplo de un dataset se representa como el vector de la i -ésima fila de su matriz correspondiente.
- Un vector puede ser visto como una matriz de una sola columna.

Matrices

- La transpuesta de una matriz A^T , corresponde a una copia de la matriz donde se intercambian las filas por las columnas.



$$A = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow A^T = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

- Podemos sumar dos matrices, siempre y cuando éstas tengan las mismas dimensiones, sumando sus elementos correspondientes: $C = A + B$ donde $C_{i,j} = A_{i,j} + B_{i,j}$. Ejemplo:

$$\begin{bmatrix} 1 & 3 \\ 1 & 0 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 7 & 5 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 1+0 & 3+0 \\ 1+7 & 0+5 \\ 1+2 & 2+1 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 8 & 5 \\ 3 & 3 \end{bmatrix}$$

Matrices

- También podemos sumar un escalar a una matriz o multiplicar una matriz por un escalar, simplemente realizando esa operación en cada elemento de una matriz: $D = a \cdot B + c$ donde $D_{i,j} = a \cdot B_{i,j} + c$

$$2 \cdot \begin{bmatrix} 10 & 6 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 2 \cdot 10 & 2 \cdot 6 \\ 2 \cdot 4 & 2 \cdot 3 \end{bmatrix}$$

- Podemos multiplicar dos matrices A y B ($C=A \cdot B$) siempre y cuando el número de columnas de A sea igual al número de filas de B. El valor de C_{ij} es igual al producto punto de la i-ésima fila de A por la j-ésima columna de B ($A_{i:} \cdot B_{:j}$).

Matrices

- Ejemplo multiplicación de matrices:

$$\begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} * \begin{bmatrix} 1 & 5 \\ 3 & 7 \end{bmatrix} = \begin{bmatrix} 3 + 12 & 15 + 28 \\ 2 + 3 & 10 + 7 \end{bmatrix}$$
$$= \begin{bmatrix} 15 & 43 \\ 5 & 17 \end{bmatrix}$$

Fuente:
<https://i1.faceprep.in/Companies-1/matrix-multiplication-in-python.png>

Matrices

- Una matriz cuadrada es una matriz que tiene el mismo número de filas y columnas.
- Una matriz cuadrada muy particular es la matriz identidad I que tiene 1s en la diagonal y ceros en todas las otras celdas.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- La inversa de una matriz A se denota como A^{-1} y cumple con la propiedad que $A \cdot A^{-1} = I$

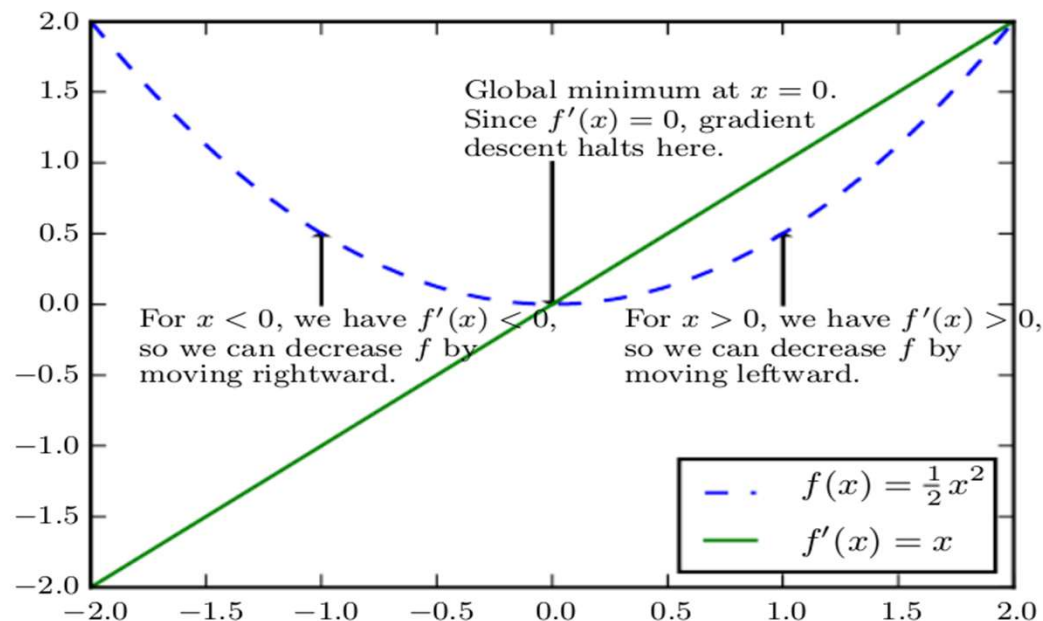
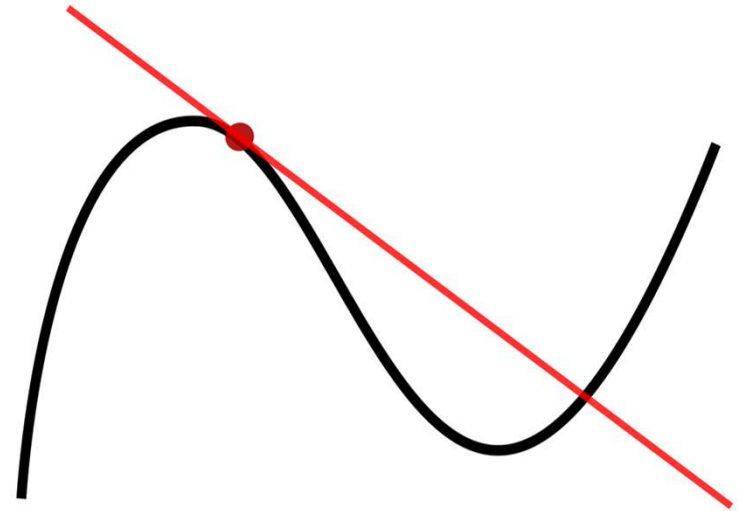
Optimización Continua

Optimización Continua

La **optimización continua** permite encontrar valores **máximo** o **mínimo** de una función matemática.

Muchos métodos de ML se plantean como problemas de optimización (ej: minimizar una función de pérdida).

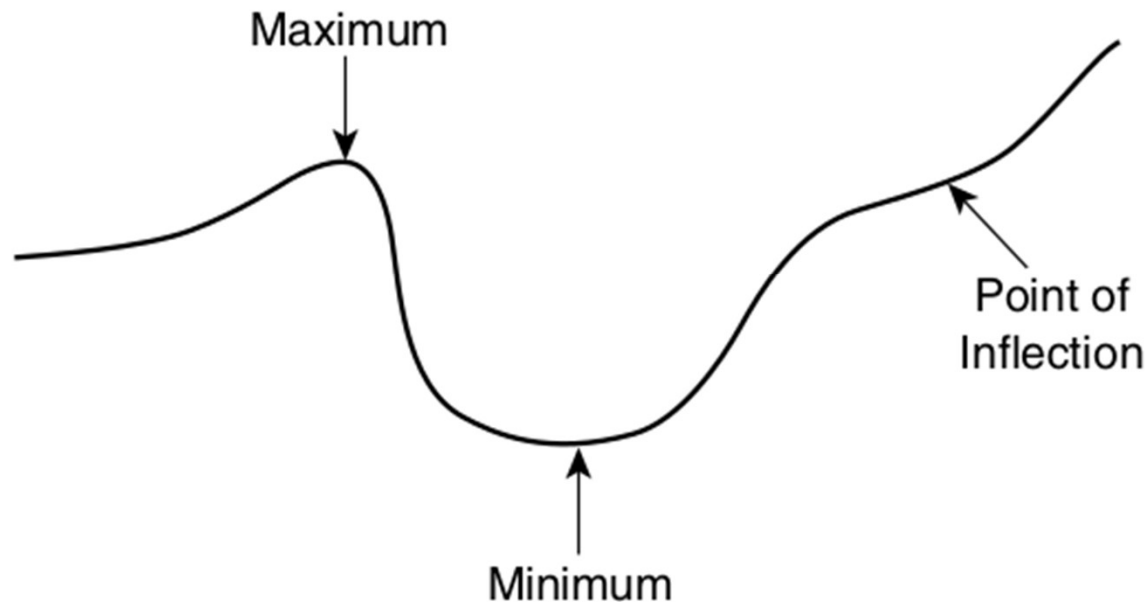
- Supongamos que tenemos una función $y = f(x)$, donde tanto x como y son números reales.
- La derivada de esta función se escribe así $f'(x)$ o así:
$$\frac{dy}{dx}$$
- La derivada $f'(x)$ nos entrega el valor de la pendiente de $f(x)$ en el punto x .



Puntos Críticos

La derivada de una función es muy útil para encontrar valores mínimos o máximos.

- Los puntos en que la derivada de una función vale cero ($f'(x)=0$) se conocen como puntos críticos: **máximo**, **mínimo** o punto de **inflexión** (o punto silla).

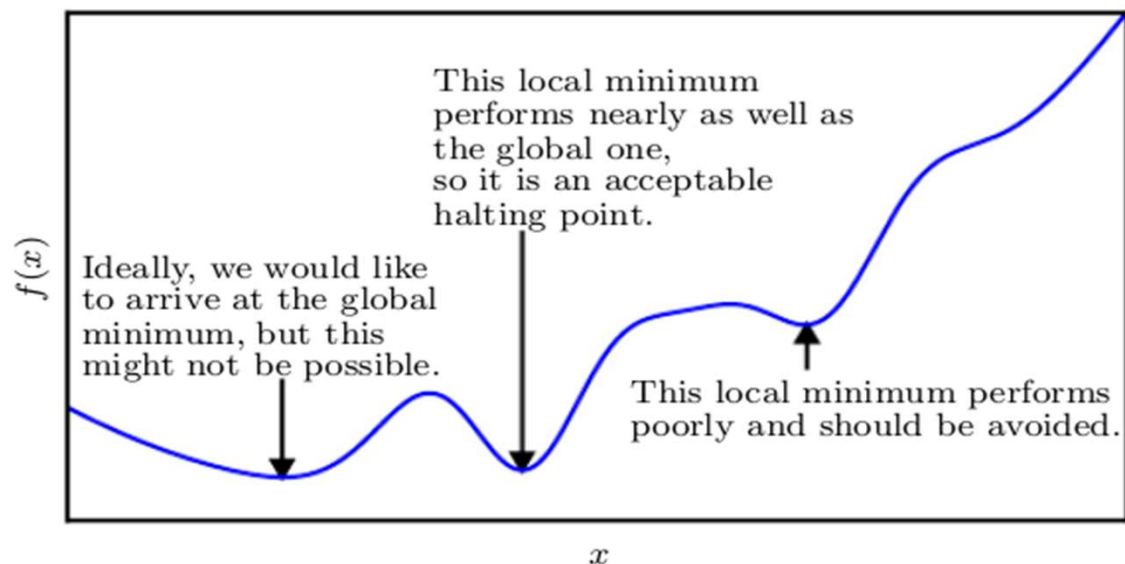


Puntos Críticos

- Para distinguir entre estos tres tipos de puntos críticos es necesario analizar la **segunda derivada** de la función $f''(x)$ (la derivada de la derivada) que nos da información sobre la curvatura de la función.

$$\frac{d^2 f}{dx^2}$$

- Otra gran dificultad al optimizar funciones es que a veces los puntos críticos pueden corresponder a mínimos o máximos **locales**.



Derivada Parcial y Gradiente

- ¿Cómo optimizamos funciones con múltiples inputs?

$$f(x_1, x_2, x_3) = 2 * x_1 + x_2^2 - 5 * x_3$$

- La **derivada parcial** $\frac{\partial}{\partial x_i} f(\mathbf{x})$ mide cómo cambia f sólo cuando hacemos un cambio en x_i .

$$\frac{\partial f}{\partial x_1} = 2, \frac{\partial f}{\partial x_2} = 2 * x_2, \frac{\partial f}{\partial x_3} = -5$$

- El **gradiente** ∇f es un vector con todas las derivadas parciales de una función.

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3} \right]$$

- Para encontrar puntos críticos en funciones de varios inputs tenemos que encontrar los valores de x donde el gradiente vale **cero**.
- Para distinguir entre máximos, mínimos y puntos silla tenemos que recurrir al **Hessiano**, que es una matriz con todas las segundas derivadas.

Optimización con restricciones

- ¿Qué pasa cuando le agregamos restricciones al problema?
- Existen dos tipos de restricciones: 1) restricciones de igualdad y 2) restricciones de desigualdad.

Restricciones de igualdad

- Supongamos que queremos encontrar el mínimo de $f(x_1, x_2, \dots, x_d)$ sujeto a las siguientes p restricciones de igualdad:

$$g_i(\mathbf{x}) = 0, \quad i = 1, 2, \dots, p.$$

- Cada restricción de igualdad es una función del tipo $g(x) = 2x_1 + 3x_2 - 3 = 0$
- Esto se puede resolver usando un método llamado **multiplicadores de Lagrange**.

Multiplicadores de Lagrange

Multiplicadores de Lagrange

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^p \lambda_i g_i(\mathbf{x})$$

1. Definimos el Lagrangiano

donde λ_i es una variable adicional llamada **multiplicador de Lagrange**.

1. Derivamos el Lagrangiano respecto a \mathbf{x} y λ , igualamos a cero y despejamos el sistema de ecuaciones.

$$\frac{\partial L}{\partial x_i} = 0, \quad \forall i = 1, 2, \dots, d$$

$$\frac{\partial L}{\partial \lambda_i} = 0, \quad \forall i = 1, 2, \dots, p.$$

La solución óptima del Lagrangiano corresponde al óptimo de $f(\mathbf{x})$ que satisface las restricciones de igualdad.

Restricciones de Desigualdad

Restricciones de desigualdad:

- ¿Qué hacemos cuando tenemos restricciones de desigualdad del tipo $h_i(\mathbf{x}) \leq 0$?
Por ejemplo $x_1 + x_2 \leq 0$.
- Se formula entonces un problema optimización con restricciones como minimizar $f(x_1, x_2, \dots, x_d)$ sujeto a las siguientes q restricciones de desigualdad:

$$h_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \dots, q.$$

- El método para resolver este problema es bastante similar al método de Lagrange descrito anteriormente.
- Sin embargo, las restricciones de la desigualdad plantean condiciones adicionales al problema de optimización.

Condición KKT

- El problema de optimización con restricciones de desigualdad se formula con el siguiente Lagrangiano:

$$L = f(\mathbf{x}) + \sum_{i=1}^q \lambda_i h_i(\mathbf{x})$$

- Una solución óptima de este problema debe satisfacer las condiciones Karush-Kuhn-Tucker (KKT):

$$\begin{aligned}\frac{\partial L}{\partial x_i} &= 0, \quad \forall i = 1, 2, \dots, d \\ h_i(\mathbf{x}) &\leq 0, \quad \forall i = 1, 2, \dots, q \\ \lambda_i &\geq 0, \quad \forall i = 1, 2, \dots, q \\ \lambda_i h_i(\mathbf{x}) &= 0, \quad \forall i = 1, 2, \dots, q.\end{aligned}$$

Condición KKT

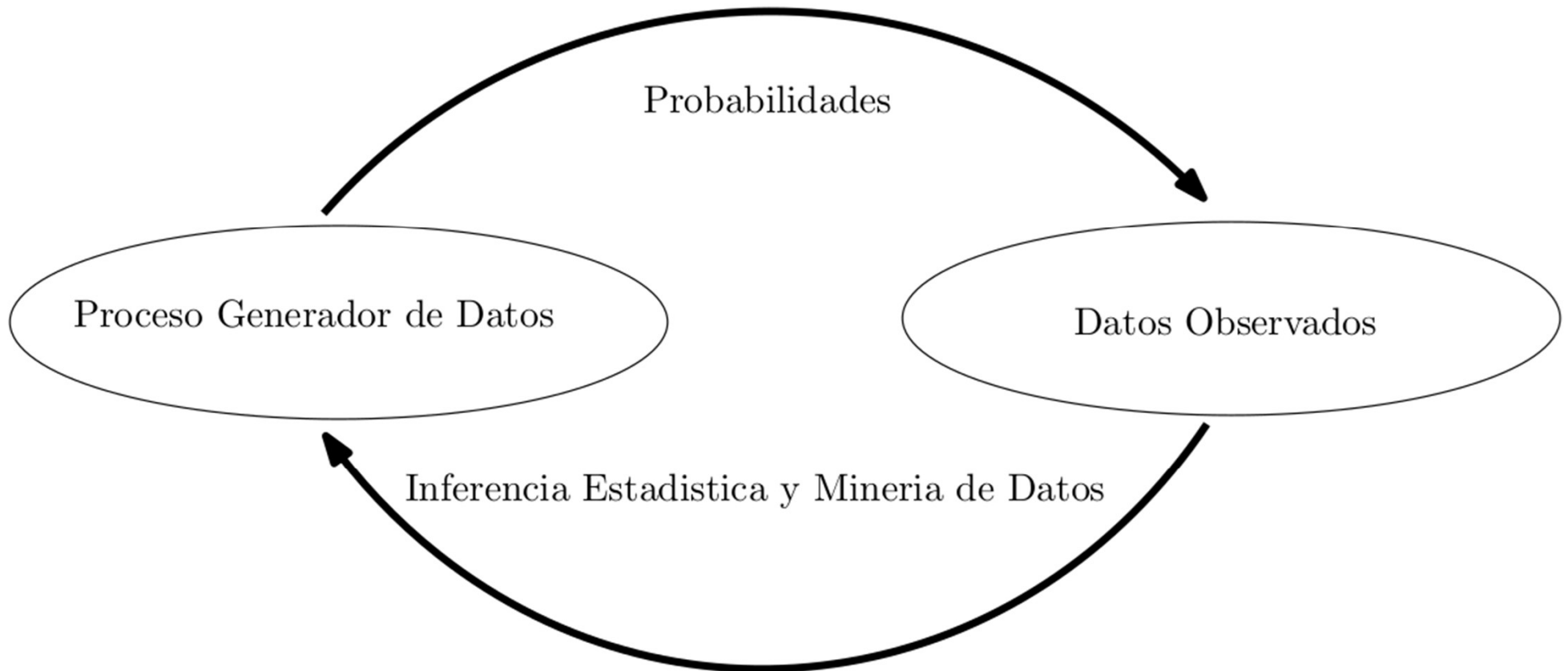
- Notemos ahora que los multiplicadores de Lagrange no pueden ser negativos.
- Las restricciones KKT nos obligan verificar que un óptimo satisfaga todas las condiciones.
- Eso puede ser muy difícil de lograr analíticamente, especialmente si se tienen muchas restricciones.
- Por lo general recurrimos a métodos numéricos como la programación lineal y la programación cuadrática.

Probabilidades

Probabilidades y Estadística

- Las probabilidades son el lenguaje de la incertidumbre que a la vez es la base de la inferencia estadística.
- El problema estudiado en probabilidades es: dado un proceso generador de datos, ¿cuáles son las propiedades de las salidas?
- El problema estudiado en inferencia estadística, minería de datos y machine learning es: dadas las salidas, ¿qué podemos decir del proceso que genera los datos observados?

Probabilidades y Estadística



Experimento Aleatorio

- Un experimento aleatorio es el acto de medir un proceso cuya salida es incierta
- El conjunto con todas las posibles salidas de un experimento aleatorio es el **espacio muestral** Ω .
- Ej: $\Omega = \{1, 2, 3, 4, 5, 6\}$ es el espacio muestral del lanzamiento de un dado.
- Un **evento** $E \subseteq \Omega$ corresponde a un subconjunto de esas salidas
- Ej: $E = \{2, 4, 6\}$ es el evento de observar un número par al lanzar un dado

Función de Probabilidad

- Una probabilidad **P** es una función de valor real definida sobre Ω que satisface las siguientes propiedades:
 - a. Para cualquier evento $E \subseteq \Omega$, $0 \leq P(E) \leq 1$
 - b. $P(\Omega) = 1$
 - c. Sean $E_1, E_2, \dots, E_k \in \Omega$ conjuntos disjuntos,

$$\mathbb{P}\left(\bigcup_{i=1}^k E_i\right) = \sum_i^k P(E_i)$$

- La probabilidad de un evento E , $P(E)$ es la fracción de veces que se observaría el evento al repetir infinitamente el experimento.

Variable Aleatoria

- Una variable aleatoria es un mapeo

$$X : \Omega \rightarrow \mathbb{R}$$

que asigna un valor real $X(e)$ a cualquier evento de Ω

Ejemplo:

- Tiramos una moneda 10 veces.
- Sea $X(\omega)$ la cantidad de caras en la secuencia de resultados.
- Si $w = CCSCCSCCSS$, entonces $X(\omega) = 6$

Variable Aleatoria

- Ejemplo: Tiramos una moneda 2 veces. Sea la variable aleatoria X la cantidad de sellos obtenidos.
- La variable aleatoria y su distribución se resume como:

e	$\mathbb{P}(e)$	$X(e)$
CC	1/4	0
CS	1/4	1
SC	1/4	1
SS	1/4	2

x	$\mathbb{P}(X = x)$
0	1/4
1	1/2
2	1/4

Distribución Acumulada

- Sea X una V.A , se define función de distribución acumulada (CDF) o $F_X : \mathbb{R} \rightarrow [0, 1]$

$$F_X(x) = P(X \leq x)$$

- Esta función nos muestra cómo se acumulan las probabilidades a medida que avanzamos en los valores de una variable aleatoria.
- Nos permite comprender la probabilidad acumulada de que un evento ocurra hasta un determinado valor.
- Por ejemplo, si tenemos una variable aleatoria que representa el tiempo que tarda un automóvil en recorrer una distancia, la función de distribución acumulada nos dirá la probabilidad acumulada de que el automóvil termine en un tiempo igual o inferior al valor dado.

Variables Aleatorias Discretas

- Una V.A X es discreta si mapea las salidas a un conjunto contable.
- Se define la función de probabilidad o función de masa de probabilidad de una V.A X discreta como $f_X(x) = P(X = x)$
- Entonces

$$f_X(x) \geq 0 \quad \forall x \in \mathbb{R} \quad \text{y} \quad \sum_i f_X(x_i) = 1$$

- La CDF de X se relaciona con f_X de la siguiente manera:

$$F_X = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} f_X(x_i)$$

Variables Aleatorias Continuas

- Una V.A X es continua si:
existe una función f_X tal que

$$f_X(x) \geq 0 \quad \forall x, \quad \int_{-\infty}^{\infty} f_X(x) dX = 1$$

y para todo $a \leq b$:

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x) dx$$

- La función f_X recibe el nombre de función densidad de probabilidad (PDF).

Variables Aleatorias Continuas

- La PDF se relaciona con la CDF como:

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$

- Luego $f_X(x) = F'_X(x)$ en todos los puntos x donde F_X es diferenciable.
- Para distribuciones continuas la probabilidad que X tome un valor particular vale siempre **cero**.

$$P(X=a)=0$$

- En VA continuas siempre tengo que calcular probabilidades para intervalos.

Variables Aleatorias Continuas

Algunas Propiedades:

1. $P(x < X \leq y) = F(y) - F(x)$
2. $P(X > x) = 1 - F(x)$
3. Si X es continua luego

$$\begin{aligned} F(b) - F(a) &= \mathbb{P}(a < X < b) = \mathbb{P}(a \leq X < b) \\ &= \mathbb{P}(a < X \leq b) = \mathbb{P}(a \leq X \leq b) \end{aligned}$$

Función Cuantil Inversa

- Sea X una V.A con CDF F , la CDF inversa o función cuantil inversa se define como:

$$F^{-1}(q) = \inf \{x : F(x) \geq q\}$$

- Para $q \in [0, 1]$ si F es estrictamente creciente y continua, $F^{-1}(q)$ es el único valor real tal que $F(x) = q$.
- Luego $F^{-1}(1/4)$ es el primer cuartil, $F^{-1}(1/2)$ la mediana (o segundo cuartil) y $F^{-1}(3/4)$ el tercer cuartil.

Algunas distribuciones

	Función de Probabilidad	Parámetros
Normal	$f_x = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$	μ, σ
Binomial	$f_x = \binom{n}{x} p^x (1-p)^{n-x}$	n, p
Poisson	$f_x = \frac{1}{x!} \lambda^x \exp^{-\lambda}$	λ
Exponencial	$f_x = \lambda \exp^{-\lambda x}$	λ
Gamma	$f_x = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp^{-\lambda x}$	λ, α
Chi-cuadrado	$f_x = \frac{1}{2^{k/2} \Gamma(k/2)} x^{(\frac{k}{2}-1)} \exp^{-x/2}$	k

Distribución Normal

- X tiene una distribución Normal o Gaussiana de parámetros μ y σ , denotado como $X \sim N(\mu, \sigma^2)$ si:

$$f_X = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$$

- Donde $\mu \in \mathbb{R}$ es el “centro” o la **media** de la distribución y $\sigma > 0$ es la **desviación estándar**.
- Cuando $\mu = 0$ y $\sigma = 1$ tenemos una **Distribución Normal Estándar** denotada por Z .
- Denotamos por $\phi(z)$ a la PDF y por $\Phi(z)$ a la CDF de una Normal estándar.
- Los valores de $\Phi(z)$, $P(Z \leq z)$ se encuentran tabulados.

Propiedades de la Normal

1. Si $X \sim N(\mu, \sigma^2)$, luego $Z = (X - \mu)/\sigma \sim N(0, 1)$
2. Si $Z \sim N(0, 1)$, luego $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$
3. Sean $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$ V.As independientes:

$$\sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$$

Ejemplo Normal

- En R podemos acceder a las PDF, CDF, función cuantil inversa y generación de números aleatorios de las distribuciones.
- Para una Normal son:

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
```

```
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

```
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
```

```
rnorm(n, mean = 0, sd = 1)
```

Ejemplo Normal

- Ejemplo: Sea $X \sim N(3, 5)$, encontrar $P(X > 1)$.

Matemáticamente esto se puede obtener con la CDF de la Normal Estandarizada:

$$\mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(Z < \frac{1-3}{\sqrt{5}}\right) = 1 - \Phi(-0,8944) = 0,81$$

Que en R se calcula así:

```
> 1-pnorm(q=(1-3)/sqrt(5))  
[1] 0.8144533
```

O directamente sobre la Normal no estandarizada:

```
> 1-pnorm(q=1,mean=3,sd=sqrt(5))  
[1] 0.8144533
```

La regla 68-95-99.7 de una Normal

- Sea X una V.A $\sim N(\mu, \sigma^2)$, se tiene:
 - a. $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0,6827$
 - b. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0,9545$
 - c. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0,9973$
- Esto se puede ver fácilmente en R para $X \sim N(0, 1)$:

```
> pnorm(1)-pnorm(-1)
```

```
[1] 0.6826895
```

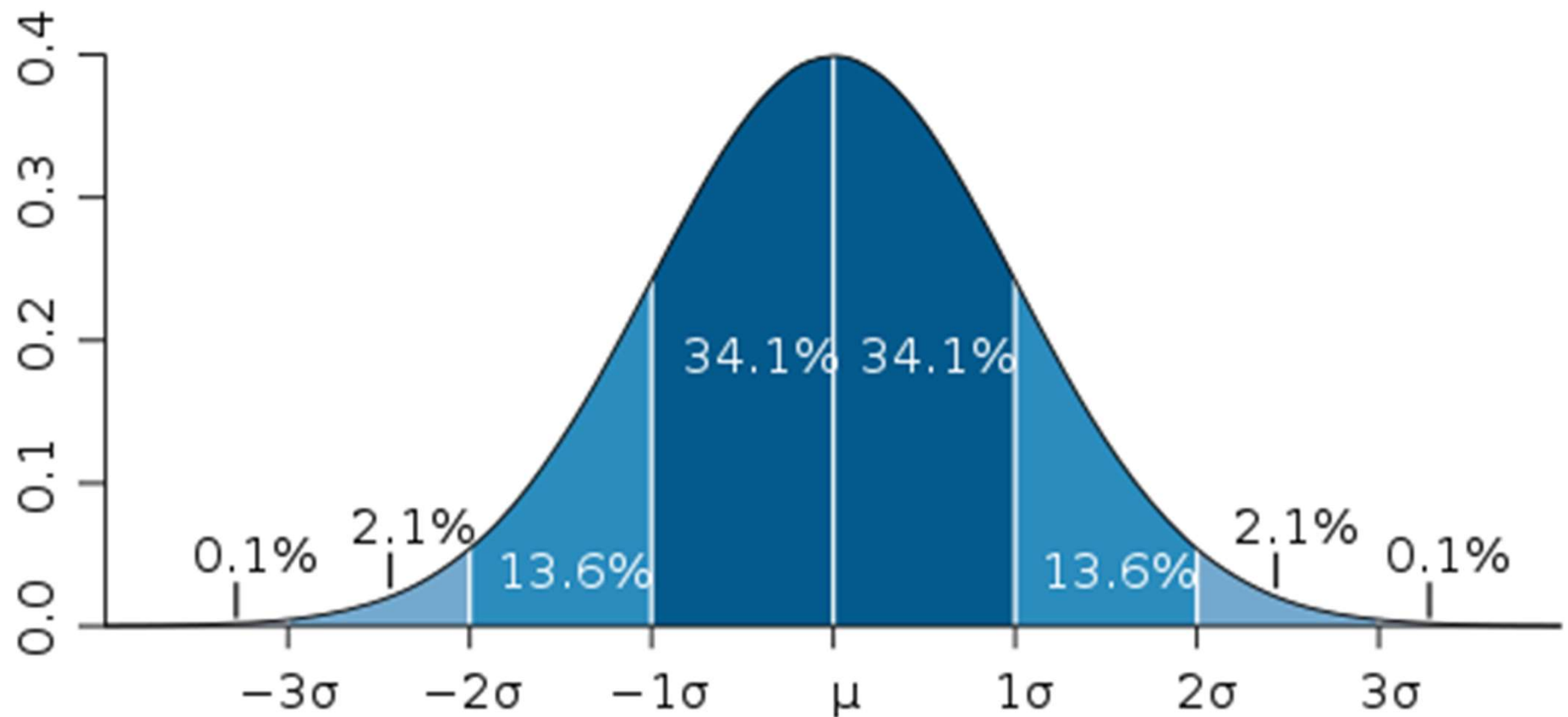
```
> pnorm(2)-pnorm(-2)
```

```
[1] 0.9544997
```

```
> pnorm(3)-pnorm(-3)
```

```
[1] 0.9973002
```


La regla 68-95-99.7 de una Normal



Simetría de la Normal

- La PDF de una normal es simétrica alrededor de μ
- Entonces $\phi(z) = \phi(-z)$
- $\Phi(z) = 1 - \Phi(-z)$

En R:

```
> dnorm(1)
```

```
[1] 0.2419707
```

```
> dnorm(-1)
```

```
[1] 0.2419707
```

```
> pnorm(0.95)
```

```
[1] 0.8289439
```

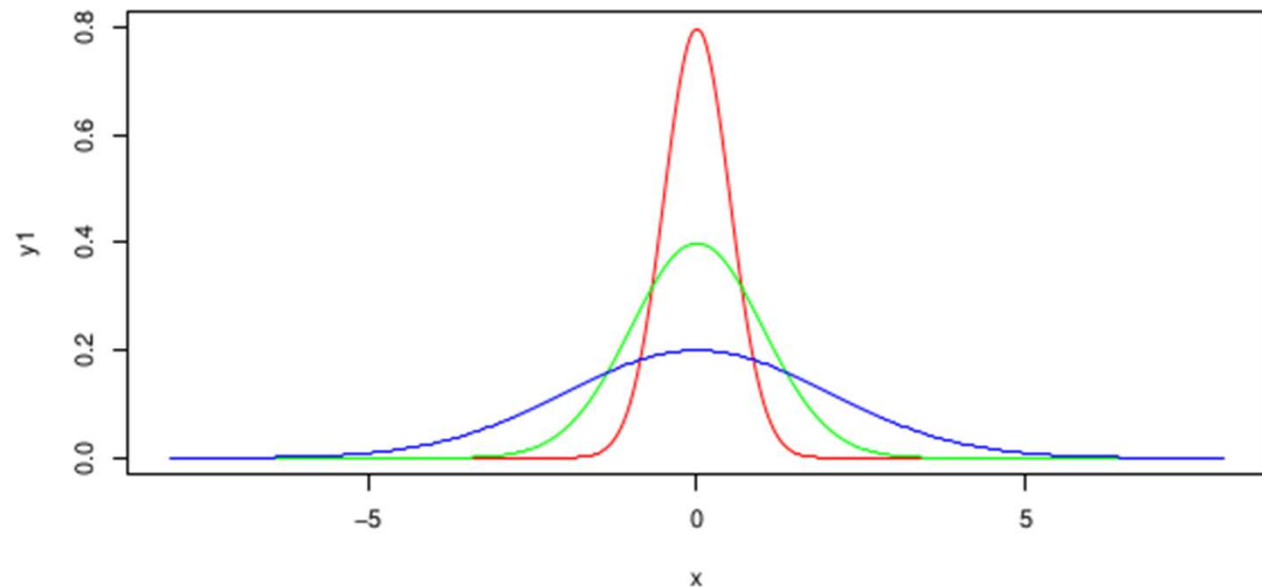
```
> 1-pnorm(-0.95)
```

```
[1] 0.8289439
```

Graficando Distribuciones Normales

Podemos graficando la PDF de Normales con distinta varianza en R con el siguiente código:

```
x=seq(-8,8,length=400)y1=dnorm(x,mean=0,sd=0.5)
y2=dnorm(x,mean=0,sd=1)
y3=dnorm(x,mean=0,sd=2)
plot(y1~x,type="l",col="red")
lines(y2~x,type="l",col="green")
lines(y3~x,type="l",col="blue")
```



Probabilidades Conjuntas y Condicionales

- La noción de función probabilidad (masa o densidad) se puede extender a más de una V.A
- Sean X Y dos V.A, $P(X, Y)$ representa la función de probabilidad conjunta.
- Las variables son independientes entre si, si:

$$P(X, Y) = P(X) \times P(Y)$$

- La probabilidad condicional para Y dado X se define como:

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X, Y)}{\mathbb{P}(X)}$$

Probabilidades Conjuntas y Condicionales

- $P(Y|X)$ se entiende como la probabilidad Y cuando yo ya sé que X está tomando un valor particular.
- Análogamente $P(Y |X)$ se puede entender como la fracción de veces que Y ocurre cuando se sabe que ocurre X .
- Luego si X e Y están relacionados, $P(Y|X)$ tiende a ser distinto a $P(Y)$.
- Si X e Y son independientes $P(Y |X) = P(Y)$.
- Ejemplo: En el experimento de lanzar un dado balanceado, sean G el evento de obtener un resultado mayor que 2 ($G = \{3, 4, 5, 6\}$) y O el evento de obtener un número impar ($O = \{1, 3, 5\}$).
- ¿Cuál es el valor de $P(G|O)$?

Probabilidades Condicionales

- Utilizando la definición de que $P(X) = (\text{casos favorables})/(\text{casos totales})$, $P(G) = 4/6$, $P(O) = 3/6$ y $P(G|O) = 2/3$.
- Observe que una vez que conocemos O , el número de casos favorables para G se redujo a $\{3, 5\}$ y el número total de casos a $\{1, 3, 5\}$.
- Ahora, según la definición anterior $P(G|O) = P(G,O)/P(O)$ donde $P(G, O) = 2/6$ (los casos favorables corresponden a la intersección entre G y O $\{3, 5\}$).
- Entonces,

$$\frac{\mathbb{P}(G,O)}{\mathbb{P}(O)} = \frac{2/6}{3/6} = 2/3.$$

Teorema de Bayes y Probabilidades Totales

- La probabilidad condicional $P(Y | X)$ y $P(X | Y)$ pueden ser expresadas en función de la otra usando el teorema de Bayes

$$\mathbb{P}(Y|X) = \frac{\mathbb{P}(X|Y)\mathbb{P}(Y)}{\mathbb{P}(X)}$$

- Probabilidades totales: sea $\{Y_1, Y_2, \dots, Y_k\}$ un conjunto de salidas mutuamente excluyentes de una V.A X , el denominador del teorema de Bayes se puede expresar como:

$$\mathbb{P}(X) = \sum_{i=1}^k \mathbb{P}(X, Y_i) = \sum_{i=1}^k \mathbb{P}(X|Y_i)\mathbb{P}(Y_i)$$

Teorema de Bayes y Probabilidades Totales

Ejemplo:

- Divido mis correos en tres categorías: A_1 =“spam”, A_2 =“baja prioridad” y A_3 =“alta prioridad”.
- Sabemos que $P(A_1) = 0.7$, $P(A_2) = 0.2$ y $P(A_3) = 0.1$, claramente $0.7 + 0.2 + 0.1 = 1$
- Sea B el evento de que el correo contenga la palabra “gratis”.
- Sabemos que $P(B|A_1) = 0.9$, $P(B|A_2) = 0.01$ y $P(B|A_3) = 0.01$ claramente $0.9 + 0.01 + 0.01 \neq 1$
- ¿Cual es la probabilidad de que sea “spam” un correo que tiene la palabra “gratis”?
- Usando Bayes y Probabilidades totales:

$$\mathbb{P}(A_1|B) = \frac{0,9 \times 0,7}{(0,9 \times 0,7) + (0,01 \times 0,2) + (0,01 \times 0,1)} = 0,995$$

Esperanza

- Sea X una V.A, se define su esperanza o momento de primer orden como:

$$\mathbb{E}(X) = \begin{cases} \sum_x (x \times f(x)) & \text{Si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} (x \times f(x)) dx & \text{Si } X \text{ es continua} \end{cases}$$

- Es el promedio ponderado de todos los posibles valores que puede tomar una variable aleatoria
- Para el caso de lanzar dos veces una moneda con X el número de caras:

$$\begin{aligned} \mathbb{E}(X) &= (0 \times f(0)) + (1 \times f(1)) + (2 \times f(2)) \\ &= (0 \times (1/4)) + (1 \times (1/2)) + (2 \times (1/4)) = 1 \end{aligned}$$

- Sean las variables aleatorias X_1, X_2, \dots, X_n y las constantes a_1, a_2, \dots, a_n

$$\mathbb{E} \left(\sum_i a_i X_i \right) = \sum_i a_i \mathbb{E}(X_i)$$

Varianza

- La varianza mide la “dispersión” de una distribución.
- Sea X una V.A de media μ , se define la varianza de X denotada como σ^2 , σ^2_X o $V(X)$ como:

$$V(X) = \mathbb{E}(X - \mu)^2 = \begin{cases} \sum_{i=1}^n f_X(x_i)(x_i - \mu)^2 & \text{Si } X \text{ es discreta} \\ \int (x - \mu)^2 f_X(x) dx & \text{Si } X \text{ es continua} \end{cases}$$

- La desviación estándar σ se define como: $\sqrt{V(X)}$

Varianza

Propiedades:

- $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \mathbb{E}(X^2) - \mu^2$

- Si a y b son constantes, luego

$$\mathbb{V}(aX + b) = a^2 \mathbb{V}(X)$$

- Si X_1, \dots, X_n son independientes y a_1, \dots, a_n son constantes, luego

$$\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i)$$

Ley de los Grandes Números

- La ley de los grandes números nos muestra que a medida que obtenemos más información o realizamos más observaciones, los resultados tienden a estabilizarse y reflejar la probabilidad esperada.
- Ejemplo: si lanzamos una moneda balanceada pocas veces es posible obtener un resultado que difiera significativamente de la probabilidad esperada.
- Por ejemplo, podríamos obtener 7 caras y 3 sellos en los primeros 10 lanzamientos.
- Luego, si aumentamos el número de lanzamientos (por ejemplo 100), los resultados se acercarán cada vez más a la probabilidad esperada (mitad de caras y sellos).

Ley de los Grandes Números

- Matemáticamente la ley de los grandes números se puede plantear de la siguiente forma (forma débil).
- Sean X_1, X_2, \dots, X_n variables aleatorias IID (independientes e idénticamente distribuidas) de media μ y varianza σ^2 .
- La VA que representa a la media o promedio se define así:

$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

- Se dice que está converge en probabilidad a μ :

$$\overline{X}_n \xrightarrow{P} \mu$$

- Esto es equivalente a decir que para todo $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|\overline{X}_n - \mu| < \epsilon) = 1$$

- Entonces la distribución de \overline{X}_n se concentra alrededor de μ cuando n crece.

Ley de los Grandes Números

- Si bien la ley de los grandes números nos dice que $\overline{X_n}$ se acerca a μ , nos dice nada sobre la distribución de esa media.
- El teorema central del límite nos permite aproximar la distribución de X_n a una distribución normal cuando n es grande.
- Eso implica que aunque no sepamos la distribución de X_i , podemos aproximar la distribución de sus medias (muestrales).
- Osea, si tomamos muchas muestras aleatorias de una variable aleatoria (de cualquier distribución) y calculamos el promedio para cada muestra, esos promedios van a distribuir como una normal.

Teorema Central del Límite

- Si bien la ley de los grandes números nos dice que $\overline{X_n}$ se acerca a μ , no nos dice nada sobre la distribución de esa media (de qué forma ocurre esa convergencia).
- El Teorema Central del Límite (TCL) nos permite aproximar la distribución de $\overline{X_n}$ a una distribución normal cuando n es grande.
- Eso implica que aunque no sepamos la distribución de X_i , podemos aproximar la distribución de sus medias (muestrales).
- Osea, si tomamos muchas muestras aleatorias de una variable aleatoria (de cualquier distribución) y calculamos el promedio para cada muestra, esos promedios van a distribuir como una normal.

Teorema Central del Límite

El Teorema Central del Límite se puede formalizar matemáticamente de la siguiente forma.

- Sean X_1, X_2, \dots, X_n variables aleatorias IID de media μ y varianza σ^2 .

- Sea
$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

- Se tiene que

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{\mathbb{V}(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow Z$$

- Lo que es equivalente a:

$$\lim_{n \rightarrow \infty} \mathbb{P}(Z_n \leq z) = \Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Teorema Central del Límite

Notaciones alternativas para el TCL:

$$Z_n \approx N(0, 1)$$

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\overline{X}_n - \mu \approx N\left(0, \frac{\sigma^2}{n}\right)$$

$$\sqrt{n}(\overline{X}_n - \mu) \approx N(0, \sigma^2)$$

$$\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0, 1)$$

Inferencia Estadística

Inferencia Estadística

- Para realizar conclusiones sobre una **población** (ej: todos los usuarios de internet), generalmente no es factible reunir todos los datos de ésta.
- Debemos realizar conclusiones razonables respecto a una población basada en la evidencia otorgada por **datos muestrales**.
- El proceso de intentar cuantificar propiedades de una población a partir de datos muestrales se conoce como **inferencia estadística**.
- A nivel muy general el objetivo de la inferencia estadística es obtener la distribución que genera los datos observados en una muestra

Inferencia Estadística

- Formalmente: dada una muestra X_1, \dots, X_n , cual es la distribución F , tal que $X_1, \dots, X_n \sim F$
- En algunos casos sólo nos interesa inferir alguna propiedad de F como su media, su varianza, etc..
- En este resumen nos enfocaremos en la **inferencia paramétrica**, la cual asume que esa distribución objetivo se puede modelar con un conjunto finito de **parámetros** $\theta = (\theta_1, \theta_2, \dots, \theta_k)$.
- Ejemplo: si asumimos que los datos vienen de una distribución normal $N(\mu, \sigma^2)$, μ y σ serían los parámetros θ .
- Un **estadístico** (muestral) es una medida cuantitativa calculada a partir de los datos.

Estimación Puntual

- La estimación puntual es el proceso de encontrar la mejor aproximación de una cantidad de interés a partir de una muestra estadística.
- La cantidad de interés puede ser: un parámetro en un modelo paramétrico, una CDF, una PDF, o una función de regresión.
- Por convención se denota a la estimación puntual del valor de interés θ como
o $\hat{\theta}$ o $\hat{\theta}_n$
- Es importante remarcar que mientras θ es un valor fijo desconocido, $\hat{\theta}$ depende de los datos y por ende es una variable aleatoria.

Estimación Puntual (2)

- Definimos a un estimador puntual de la siguiente manera.
- Sean X_1, \dots, X_n n observaciones IID de una distribución F .
- Un estimador puntual $\hat{\theta}_n$ de un parámetro θ es una función de X_1, \dots, X_n :

$$\hat{\theta}_n = g(X_1, \dots, X_n)$$

- El sesgo (bias) de un estimador se define como:

$$\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$

- Un estimador es insesgado si:

$$\mathbb{E}(\hat{\theta}_n) = \theta \text{ o } \text{bias}(\hat{\theta}_n) = 0$$

Estimación Puntual (3)

- La distribución de $\hat{\theta}_n$ se conoce como la **distribución muestral** del estimador.
- Representa la distribución que tendría el estimador si tomáramos todas las muestras posibles de tamaño n .
- Ejemplo: imaginemos que nuestro estimador $\hat{\theta}_n$ es la media muestral \overline{X}_n

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

- Si nuestra población tiene tamaño N y nuestras muestras son de tamaño n , la distribución muestral corresponde a las frecuencias relativas de todas las medias obtenidas para todas las muestras posibles de tamaño n sobre nuestra población.

Estimación Puntual (4)

- La desviación estándar de $\hat{\theta}_n$ **se** (que a su vez proviene de la distribución muestral) conoce como error estándar se:

$$se(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}$$

- El error estándar nos habla sobre la variabilidad del estimador entre todas las posibles muestras de un mismo tamaño.
- Para el ejemplo anterior sería la variabilidad de todas las medias muestrales posibles.

Estimación Puntual (5)

- Sea X_1, X_2, \dots, X_n una muestra aleatoria de una población de media μ y varianza σ^2 .
- Sea nuestro estimador $\hat{\mu}$ la media muestral \overline{X}_n

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

- Veamos que $\hat{\mu}$ es un estimador insesgado de la media poblacional μ :

$$\mathbb{E}(\overline{X}_n) = \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \times \mathbb{E}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n}(n \times \mu) = \mu$$

Estimación Puntual (6)

- El error estándar de nuestro estimador $se(\bar{X}_n) = \sqrt{\mathbb{V}(\bar{X}_n)}$ se calcula como:

$$\mathbb{V}(\bar{X}_n) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \mathbb{V}\left(\sum_{i=1}^n X_i\right) = \frac{n}{n^2} \mathbb{V}(X_i) = \frac{\sigma^2}{n}$$

- Entonces:

$$se(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

- Esto nos dice que la variabilidad de este estimador decrece con el tamaño de la muestra en un orden de raíz cuadrática.

Estimación Puntual (7)

- Nuestra estimación anterior del error estándar requiere conocer el σ de la población.
- En escenarios realistas este valor es desconocido y también debemos estimarlo.
- Cuando queremos estimar la varianza de una población a partir de una muestra hablamos de la **varianza muestral**.

Estimación Puntual (8)

- Existen dos estimadores comunes de este parámetro poblacional:

a. La versión sesgada:

$$s_n^2 = \frac{1}{n} \sum_i^n (X_i - \bar{X}_n)^2$$

b. La versión sin sesgo:

$$s^2 = \frac{1}{n-1} \sum_i^n (X_i - \bar{X}_n)^2$$

- Entonces el error estándar de la media muestral como estimador de la media cuando no conocemos la varianza poblacional se calcula así:

$$\hat{se}(\bar{X}_n) = \frac{s}{\sqrt{n}}$$

Estimación Puntual (9)

- Una característica deseable de un estimador es que sea insesgado y de mínima varianza (error estándar).
- Decimos que un estimador puntual $\hat{\theta}_n$ de un parámetro θ es **consistente** si converge al valor verdadero cuando el número de datos de la muestra tiende a infinito.
- Por ejemplo, para la media muestral $\mathbb{E}(\bar{X}_n) = \mu$ lo que implica que el bias = 0 y que tiende a cero cuando $n \rightarrow \infty$.
 $se(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$
- Entonces \bar{X}_n es un estimador consistente de la media.

Intervalo de Confianza

- ⌋ Sabemos que el valor de un estimador puntual varía entre una muestra y otra
- ⌋ Es más razonable encontrar un intervalo donde tengamos garantías que el valor real del parámetro se encuentra dentro del intervalo con una cierta frecuencia.
- ⌋ La forma general de un intervalo de confianza es la siguiente:

$$\text{Intervalo de Confianza} = \text{Estadístico Muestral} \pm \text{Margen de Error}$$

- ⌋ Entre más ancho el intervalo mayor incertidumbre existe sobre el valor del parámetro.

Intervalo de Confianza (2)

- Se define un intervalo de confianza para un parámetro poblacional desconocido θ con un nivel de confianza $1 - \alpha$, como el intervalo C_n = (a, b) donde:

$$P(\theta \in C_n) = 1 - \alpha$$

- Donde $a = a(X_1, \dots, X_n)$ y $b = b(X_1, \dots, X_n)$ son funciones de los datos.
- El valor α se conoce como el nivel de significancia, generalmente se toma como 0,05 lo que equivale a trabajar con un nivel de confianza de 95 %.
- La significación estadística ayuda a cuantificar si un resultado se debe probablemente al azar o a algún factor de interés.
- Básicamente, un resultado estadísticamente significativo es un resultado que no es atribuible al azar.

Intervalo de Confianza (3)

- Existe mucha confusión de cómo interpretar un intervalo de confianza
- Una forma de interpretarlos es decir que si repetimos un mismo experimento muchas veces, el intervalo contendrá el valor del parámetro el $(1 - \alpha) \%$ de las veces.
- Esta interpretación es correcta, pero rara vez repetimos un mismo experimento varias veces.
- Una interpretación mejor:
 - un día recolecto datos creo un intervalo de 95 % de confianza para un parámetro θ_1 .
 - Luego, en el día 2 hago lo mismo para un parámetro θ_2 y así reiteradamente n veces.
 - El 95 % de mis intervalos contendrá los valores reales de los parámetros.

Intervalo de Confianza (4)

- Se tienen n observaciones independientes X_1, \dots, X_n IID de distribución $N(\mu, \sigma^2)$
- Supongamos que μ es **desconocido** pero σ^2 es **conocido**.
- Sabemos que \bar{X}_n es un estimador insesgado de μ
- Por la ley de los grandes números sabemos que la distribución de \bar{X}_n concentra alrededor de μ cuando n es grande.
- Por el CLT sabemos que \bar{X}_n converge a una distribución muestral Normal cuando n es grande:

$$Z = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

